CI Experimental Design Workshop Manual

"There's a flaw in your experimental design.
All the mice are scorpios."

November 2014

Bioinformatics & Genomics Core Facilities
Cancer Research UK Cambridge Institute
University of Cambridge

Authors:
Roslin Russell (roslin.russell@cruk.cam.ac.uk)
Sarah Dawson (sarah.dawson@cruk.cam.ac.uk)
Sarah Vowler (sarah.vowler@cruk.cam.ac.uk)
Sarah Leigh-Brown (sarah.leigh-brown@cruk.cam.ac.uk)

CONTENT

The Importance of Good Experimental Design

Fisher's Fundamental Experimental Design Principles

Ronald A Fisher (1890-1962) was a great statistician and quantitative geneticist, and one of the pioneers in good experimental design.

Fisher stated the following:

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of." (1938).

It can take a lot of time and effort to conduct a 'post-mortem' of the experimental results, sometimes even longer than the analysis for the same experiment had it worked. Your results are less likely to become a 'post-mortem' if you apply the principles of good experimental design**,** as this will increase the validity of your experiment. If a study is valid then it truly represents the population it was intended to represent.

This is why you must plan your experiment before you start on any data/sample collection and processing. If possible, planning should involve an Experimental Design meeting at the very beginning of the experiment and continued contact with your Bioinformatics Analyst as the design evolves.

Without a valid design, valid scientific conclusions cannot be drawn

Fisher's principles of scientific inference underlie the design and analysis of experiments to investigate the causal effects of treatments on a response variable of interest. Fisher stated clearly that an understanding of these principles is essential to everyone engaged in science.

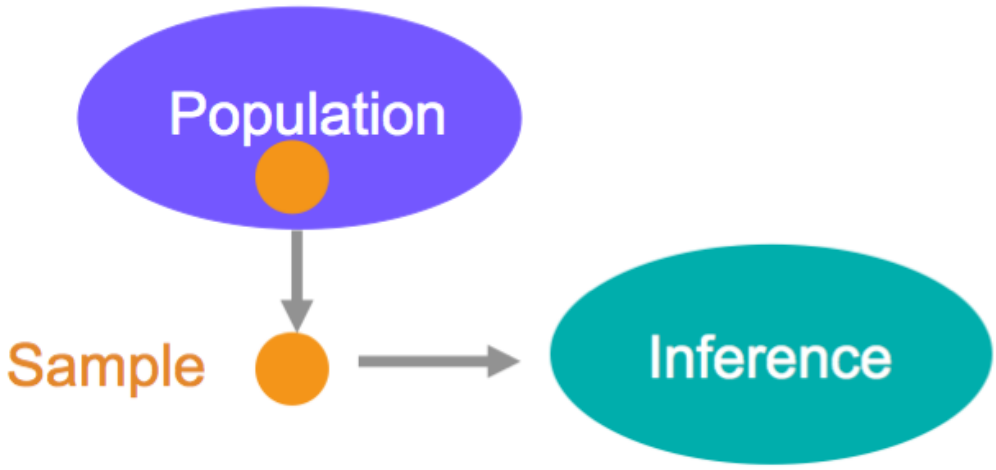
Three fundamental experimental design principles are attributed to Fisher:

- **Replication**: Having more samples, in general, reduces variability, giving you a more precise answer to your experimental question.

- **Randomisation**: Assigning samples to treatments randomly helps to protect from bias in results, so your results are more reliable.

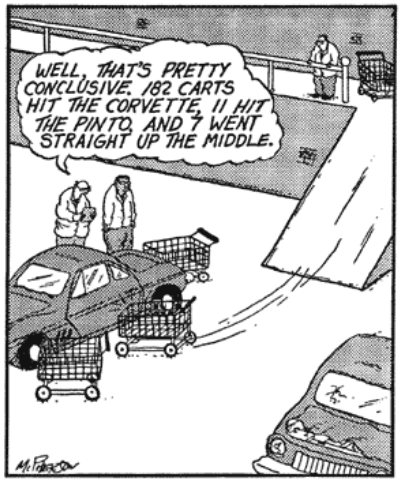
- **Blocking:** Using a homogenous group of experimental units e.g. a litter or cell line passage allow your results to be more reproducible.

Generalizability & Inference



When we carry out an experiment we have a group that we intend our results to apply to, this group is known as the population. For example we may wish results to apply to all women in the UK with breast cancer. Taking a random selection from the population to be included within your experiment should ensure that your sample is representative of the population from which it is drawn. This means that results that you obtain will be generalizable to the population of interest. When we carry out an analysis of the data obtained from this population we obtain inferential statistics. Inferential because we are inferring what will happen in the whole population from the sample selected.

Experimental Validity



Researchers at MIT prove that rolling shopping carts will almost invariably hit the most expensive car in their vicinity.

When we ask whether a piece of research is valid or not we are essentially asking, "is this true?" Answering this question will inevitably involve a degree of subjective judgement, but by managing the various threats to the validity of our research we can improve our chances of producing valid work. As we will discuss later in the manual, the main threats to validity are chance, bias and confounding**.**

Chance: the measurements we make while doing research are nearly always subject to *random* variation. Determining whether findings are due to chance is a key feature of statistical analysis. As we will discuss further later, the best way to avoid error due to random variation is to ensure your sample size is adequate.

Bias: whereas chance is caused by *random* variation, bias is caused by *systematic* variation. A systematic error in the way we select our patients, measure our outcomes, or analyse our data will lead to results that are inaccurate.

Confounding: this is similar to bias and is often confused. However, whereas bias involves error in the measurement of a variable, confounding involves error in the interpretation of what may be an accurate measurement. A classic example of confounding is to interpret the finding that people who carry matches are more likely to develop lung cancer as evidence of an association between carrying matches and lung cancer. Smoking is the confounding factor in this relationship as smokers are more likely to carry matches and they are also more likely to develop lung cancer.

Experimental validity is broken down into two categories:

- **Internal Validity**: Is a study able to determine if a causal relationship exists between the treatment (independent variables) and the response (dependent variables)? In other words, can we be reasonably sure that the observed change was caused by the treatment? An experiment has high internal validity if it has a high probability of getting the correct answer.

- **External Validity**: Is the study generalizable? Can we be sure that the results of our study truly represent the entire population? An experiment will have high external validity if the results can be generalised to other conditions or situations.

A research finding may be entirely valid in one setting but not in another. For example, an experiment using only a single strain of mice may have high internal validity, but if the same results are not seen with other strains of mice, then it will have low external validity.

It is sometimes acceptable to do an experiment with high internal validity but no exploration of its external validity, provided it is made clear that the external validity is unknown. However, increasingly journals are insisting that external validation is provided in the paper, particularly in a medical setting. It may be difficult to get your papers published without external validation.

Note that a result cannot have high external validity unless it first has high internal validity. A study that is valid but not generalizable is at least useful in the setting in which it was carried out. However, there is no point trying to use or generalise the findings of an invalid study.

You should aim to maximise both validity and generalizability, but sometimes you must make a trade-off between the two, for example:

- A broad patient selection will improve generalizability but may impair validity if inappropriate patients are included.

- Undertaking a study in a well funded, specialist centre may improve validity but will undermine generalizability.

- The results of a multicentre study will be more generalizable than a single centre study. However, trial procedures that reduce bias (blinding, outcome measurement and follow up) may be easier to control in a single centre.

- With random selection, increasing the sample size will enhance generalisability and reduce the risk of a random error rendering the results invalid, but trial procedures may be more difficult to control.

Why is this important for your research?

Compared with Fisher's day, large-scale biological experiments today differ in that we are able to simultaneously measure thousands of response dependent variables (e.g. expression levels) for each experimental unit (e.g. mouse) rather than only one or a few. As a result, the principles of experimental design are more important now than ever.

A poorly designed experiment can be costly to the individual investigator, both in terms of resources and reputation:

What do you have to lose?

- **Money**: kits, reagents and services used during your experiment may be entirely wasted and this can be very costly. For example, the cost of each mouse in an experiment is around £250. The cost of one lane of sequencing which you do not use for analysis is £1350. You could waste precious grant money or the building's hard-earned CRUK funding.
- **Time:** The time you spend collecting and preparing samples for a poorly designed experiment is wasted.
- **Samples:** If you are working with patient tissue samples they normally provide limited material and are therefore and precious and irreplaceable.
- **Reputation:** If your experimental results are not valid or reproducible this can lead to widespread reputational damage, and retraction publications reporting poorly designed experiments.
- **Scientific progress:** It is a pre-requisite for most journals to submit microarray or Next Generation Sequencing data to public repositories when you submit a paper. This is so others can re-analyse your results for new biological questions. Poorly designed experiments hinder others trying to use meta-analysis across multiple data sets. It can take analysis from multiple external groups before errors in the original data source are revealed.
- **Ethical issues:** Use of animals in research is tightly regulated, and an experiment using animals which is poorly designed risks wasting those animals.

MIBBI

The Minimum Information for Biological and Biomedical Investigations (MIBBI) project specifies the information required for research papers that use a variety of different technologies:

http://mibbi.sourceforge.net/portal.shtml

Ethical Concerns in Animal Studies

Replacement, Refinement, Reduction (The 3Rs)

Animal experiments are necessary in cancer research, but they need to be appropriately regulated, and here in the UK we have some of the tightest regulation in the world.

If you are working with animals in your research you are legally required to attend the Home Office Licencing training course, which can be organised through the University of Cambridge.

The National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) issues guidelines for use of animals in research:

- **Replace**: Find innovative solutions to replace animals in research with non-animal alternatives.
- **Reduce:** the number of animals used in experiments.
- **Refine**: Minimise suffering and improve animal welfare by refining husbandry and procedures.

Use of Animals in Research

"For scientific, ethical and economic reasons, experiments involving animals should be appropriately designed, correctly analysed and transparently reported. This increases the scientific validity of the results and maximizes the knowledge gained from each experiment."

Kilkenny et al. PlosOne, 2009

The NC3Rs carried out a survey of the quality of reporting, experimental design and statistical analysis of recently published research using laboratory animals (Kilkenny et al, 2009). They analysed 271 papers from 1999-2005, covering a representative sample

of biomedical research and an extensive range of journals across the impact factor spectrum, including Nature and Science. They reported:

- 4% of papers did not mention how many animals were used in the experiment anywhere in the paper.
- Of the studies that did say how many were used, none explained why they had chosen their particular number of animals.
- 35% of the papers gave one figure for the number of animals used in the methods, and then a different number of animals appeared in the results.
- Only 8% reported the raw data so you could repeat their analysis.
- About half the studies did not report the numbers of animals in each group in their tables.
- Only 12% of the animal studies used randomisation[1] and only 14% used blinding[2] (both are discussed in more detail later in the manual).

The ARRIVE guidelines

The Animal Research: Reporting *In Vivo* Experiments (ARRIVE) guidelines have been developed by the NC3Rs to improve standards of reporting and ensure that the data from animal experiments can be fully evaluated and utilised (Kilkenny et al, 2010). The guidelines are primarily aimed at scientists writing up their research for publication and for those who are involved in peer review.

What can you learn from this?

The NC3Rs findings are shocking and they expose how poor the execution and reporting of experimental design can be in animal research.

The most important thing about good science is to be clear about the shortcomings of your own method**.** Every scientific experiment has to take short cuts for practical reasons, cost or ethics – but you have to be clear about those shortcomings and report your findings with caveats and explain why they are potential sources of bias in your results.

Whether we are working with animals, cell-lines or patient samples, we should follow best practices at every stage of the experiment, from when we plan, right up to when we report and publish our findings. The Core Facilities can help you achieve this, please feel free to contact us at any stage of your experiment.

There are regular Experimental Design Meetings on Tuesdays and Statistics Clinics on Wednesdays:

Experimental Design meetings: CRIExperimentalDesign@cruk.cam.ac.uk
Statistics Clinic: CRIStatsClinic@cruk.cam.ac.uk

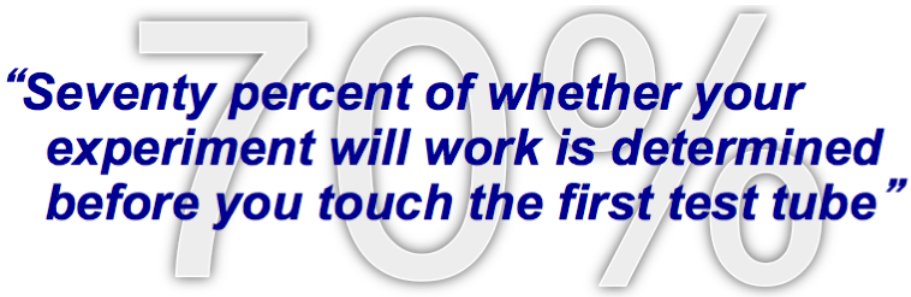
[1] Randomisation is chance assignment to each group.
[2] Blinding is where the subject and/or experimenter and/or statistician are unaware of group assignment and depending on the experiment this may or may not be possible.

Resources for use of Animals in Research:

NC3Rs website: http://www.nc3rs.org.uk/
NC3Rs short course http://www.3rs-reduction.co.uk
ARRIVE Guidelines: http://www.nc3rs.org.uk/page.asp?id=1357
CRUK NCWO & NCCO: Tony Davidge Tony.Davidge@cruk.cam.ac.uk

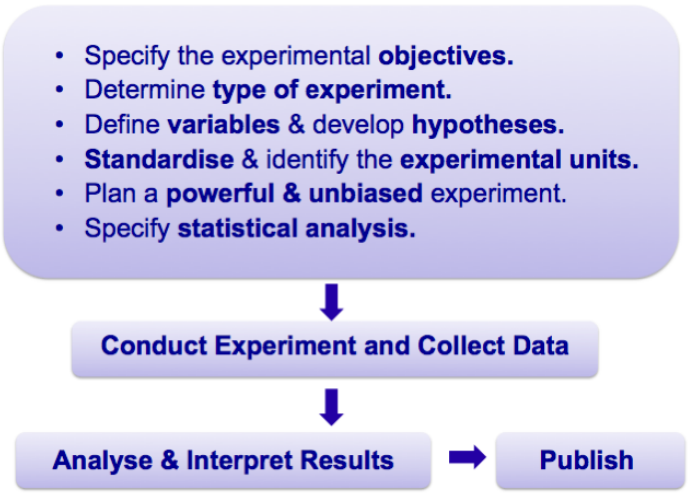
Planning

Rushing into experiments without strategic planning invites failure.



Tung-Tien Sun (2004).

Experimental design requires strategic planning in advance of starting the experiment because the validity of the experimental result depends on it. This includes planning of how to use the available expertise, time and money, choice of material, reagents, methodology (in vivo, in vitro and in silico) and technology.



While planning your experiment, it is important to include members of the Core Facilities who are involved in processing your samples (e.g. Genomics), and involve a Bioinformatics Analyst and/or Statistician. Together, they can help you decide on the appropriate experimental design and methods for the data analysis.

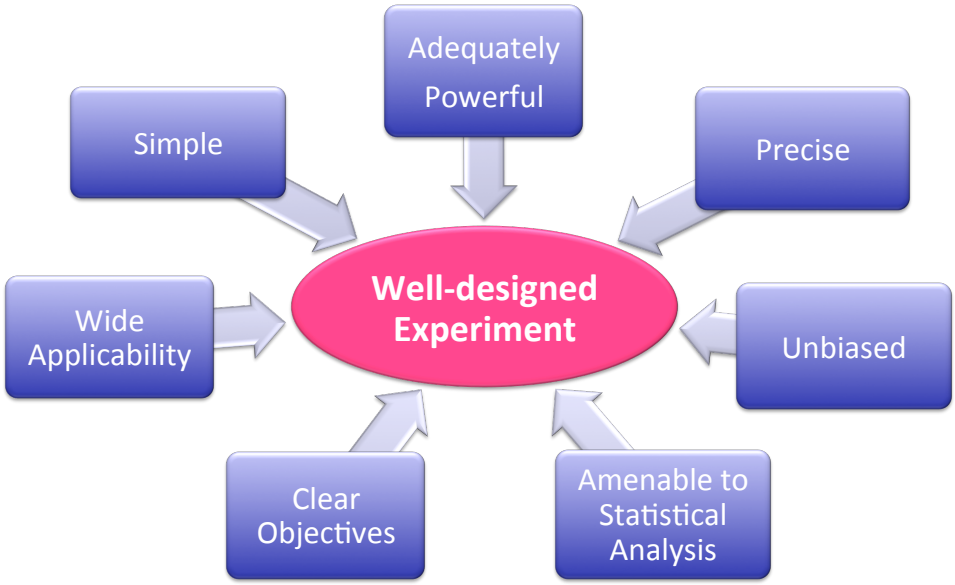
Strategic planning leads to 'best' scientific practice**.** Remember to plan **before** you collect data / process samples. You and others are going to make research decisions based on the results (e.g. choosing the next experiment).

A summary of what generally needs to be considered during the planning process is shown in the figure below. All the points shown in the diagram will be discussed in more detail later in the manual. Once you are clear on what your objectives are, you can design an experiment which is:

Powerful: High probability of detecting a treatment effect should it exist.

Unbiased: no group of samples will be treated differently from any other (except for the applied interventions), as this may cause bias.

A Well-Designed Experiment



Clear Objectives

Planning involves a lot of information gathering. With any experiment, it is important to start with clearly stated objectives and to think about the following:

- What are the motives for doing the experiment?
- What are your experimental aims?
- What questions are you asking?
- Are some questions more important than others?
- What type of experiment are you considering?
- Are you going to do a feasibility experiment to test logistics?
- Are you going to do a pilot experiment to gain preliminary information?
- Are you going to test a hypothesis about the effect of a treatment?
- Are you exploring the effect of some treatment?
- Are you estimating parameters such as a dose-response relationship or group means?

It is also important to consider what you are measuring and how you are going to make these measurements.

- Are you measuring, for example, changes in expression or tumour growth?
- Will you be using real-time PCR, microarrays or Next Generation Sequencing to measure gene expression?
- Are you interested in differential expression at the gene level or splice isoform level?

You may find it helpful to think about what you are measuring mathematically. For example: are you comparing the means of two groups, proportions or variance, or are you interested in studying a relationship (a trend)?

Before the experiment is started the type of data to be collected should be identified. Then a statistical analysis plan should be written detailing exactly how the analysis will be carried out. Anyone starting an experiment should know in general terms how the resulting observations are to be treated. For many expensive 'omic' technologies such as microarrays, Next Generation Sequencing and stable isotope protein labelling (e.g. SILAC, iTRAQ and TMT), expert advice is required in the analysis of the data generated. It is important to work with the person who will analyse your data at the planning stage to ensure that the experimental design is appropriate and optimal for these technologies.

Wide Range of Applicability

Having a wide range of applicability in your experiment addresses secondary questions. It is often desirable to plan your experiment so the results are applicable over a wide range of conditions. For example you might include mice of both sexes, several strains, different ages, different drugs, environments and prior treatments. The range of applicability is explored using factorial and randomised block designs (RBDs), which can sample different situations.

A factorial design involves two or more factors (independent variables) in a single experiment. These designs are classified by the number of levels of each factor and the number of factors. For example, gender has two levels, male and female. A 2x2 factorial will have two factors, each with two levels, and a 2x2x2 factorial will have three factors each with two levels.

Typically, there are many factors such as gender, genotype, diet, experimental protocols and age, which can influence the outcome of an experiment. These often need to be investigated in order to determine the generalisability of a response. It may be important to know whether a response is only seen in, say, females but not males. One way to do this would be to do separate experiments in each sex. This "OVAT" or "One Variable at A Time" approach is, however, very wasteful of scientific resources. A much better alternative is to include both sexes or more than one strain, etc, in a single factorial experiment rather than using several sequential experiments. Such designs can include several factors without using excessive numbers of experimental subjects. However, it is much more important to make sure that these experiments are designed correctly, as more results are at stake.

Factorial designs are efficient and provide extra information about the interactions between the factors, which cannot be obtained when using single factor designs. There are alternative approaches to factorial designs and we recommend that you speak to a statistician for further information if you are interested in looking at interaction effects. An interaction effect tells us about the influence of one independent variable on another. We can also assess what effects multiple independent variables have on a single dependent variable. Fisher, RA (1960) highly recommends the use of factorial designs for increasing the precision and scope of the experimental investigation without increasing the number of samples:

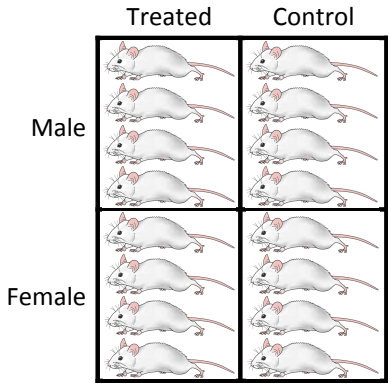
"If the investigator confines his attention to any single factor we may infer either that he is the unfortunate victim of a doctrinaire theory as to how experimentation should proceed, or that the time, material or equipment at his disposal is too limited to allow him to give attention to more than one aspect of his problem.....

... [by using factorial designs] an experimental investigation, at the same time as it is made more comprehensive, may also be made more efficient if by more efficient we mean that more knowledge and a higher degree of precision are obtainable by the same number of observations."

In vivo example:

Assuming that the animal is the experimental unit, the experiment in this example has two independent factors treatment and gender (see diagram).

The aim is usually to see whether the two factors are independent. This is a 2x2 factorial design because there are two factors each at two levels: gender factor (male and female levels) and treatment factor (treated and control levels).



Treatment: Control versus Treated, represented by the two columns.

Gender: Male versus Female, represented by the rows.

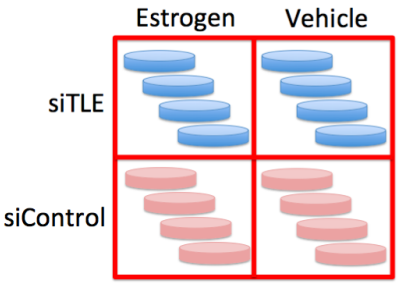
Factorial designs are powerful because differences among the levels of each factor are determined by averaging across all other factors. If columns in the figure on the left represent "Treated" and "Control" and the means are estimated by averaging across the two columns, which might represent males and females. This assumes that the males and females respond in the same way to the treatment, an assumption that is tested in the statistical analysis using a two-way analysis of variance with an interaction term, assuming that such an analysis is appropriate.

If the two sexes do not respond in the same way then this is known as an interaction and the differences will need to be looked at separately for each sex.

This hypothetical example requires several tests to prove statistically significant differences as you are making several comparisons:

Effect	Test
Main effect of treatment	Difference in treated vs. control in males
	Difference in treated vs. control in females
Main effect of gender	Difference in males vs. females in treated mice
	Difference in males vs. females in control mice
Interaction of treatment and gender	The difference of differences: the difference of treatment in males against the difference of treatment in females.

In vitro example:



In this 2x2 factorial design your goal is to confirm whether TLE1 is dependent on active estrogen receptor (ER). Therefore you measure the global expression values using microarrays, with estrogen and in a vehicle control (without estrogen) in the siTLE1 cells, and also in the siControl cells.

What you may find with the differential gene expression analysis is that certain genes are statistically significant when comparing the 'estrogen versus vehicle treated' groups for the siTLE, but these genes are not found to be statistically different when comparing the same treatment groups in the siControl cells.

There is an issue here with the analysis. You can say there is a statistically significant effect for your 'estrogen versus vehicle treated' siTLE cells, and you can say there is no evidence of an effect in the siControl cells. However, you cannot say siTLE and siControl cells respond to estrogen differently. To say this, you would have to do another type of statistical test to compare the difference in differences, i.e. the difference between the estrogen-induced change in the siTLE cells against the estrogen-induced change in the siControl cells. In other words, you would test the interaction between cell line and estrogen treatment.

More factorial design examples

http://www.3rs-reduction.co.uk/html/10__factorial_experiments.html

Factorial Designs and Incorrect Analysis

Unfortunately, although such designs are widely used, they are often incorrectly analysed, as recently presented in a survey by Niewenhuis et al. (2011):

Number of studies	513
Factorial designs	153 (30%)
Correctly analysed	78 (50%)

See **Ben Goldacre's** blog on this:
http://www.guardian.co.uk/commentisfree/2011/sep/09/bad-science-research-error

and another interesting blog on the **top 5 statistical 'faux pas'**:
http://www.methodspace.com/profiles/blogs/top-5-statistical-fax-pas

Keep it Simple

Experiments should not be so complicated that mistakes are made in their execution, or the statistical analysis becomes unduly complicated.

For example, if you need to process a lot of samples at one time, think about whether it is practical to do this without hurrying and making any errors. It may be easier to split the samples into batches over several days. However, careful consideration should be given to sample allocation. For example, a randomised block design (see later in the manual) would control any bias introduced due to processing in batches and the analysis would be able to test for a 'batch effect'.

Clearly written protocols and standard operating procedures (SOPs) should be used. In some cases it may be necessary to work to Good Laboratory Practice standards (http://en.wikipedia.org/wiki/Good_Laboratory_Practice).

In certain situations, feasibility or pilot studies should be used before starting a major experiment to ensure that the experiment is logistically efficient and to give some preliminary indication of likely results (which may be used to base a sample size calculation on for the major part of your experiment). A pilot is a small experiment designed to test logistics and gather information prior to a larger study, in order to improve the larger studies quality and efficiency. A pilot study can reveal deficiencies in the design of a proposed experiment or procedure and these can then be addressed before time and resources are expended on large-scale studies. For further information, see the NC3Rs documentation on conducting pilot studies. https://www.nc3rs.org.uk/conducting-pilot-study.

Adequately Powerful

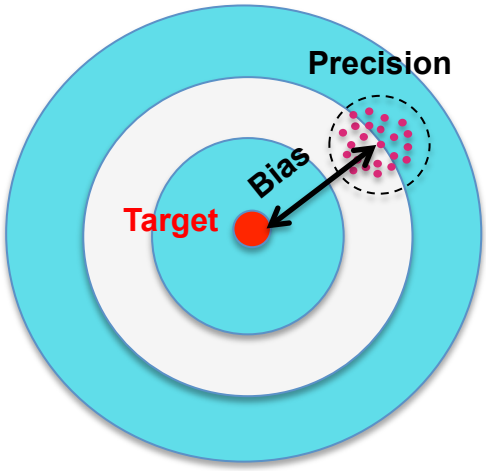
A well-designed study should be adequately powerful: it should have a high probability of detecting an effect of clinical or scientific importance if it is present. Adequate power is achieved by:

- Using appropriate numbers of subjects (sample size)
- Controlling inter-subject variation (e.g. using randomisation)

Experiments that lack power will give you too many false negative results. i.e. they may not detect important effects that truly exist. Power will be discussed in more detail late in the manual – section on 'Replication Sample Size & Power'.

Precise

Precision (or variation) is the amount of scatter around the true value. It refers to the degree of agreement for a series of measurements.



A well-designed experiment will give acceptable precision and allow you to estimate the quantity of interest. Precision is increased by using appropriate instrumentation, meticulous laboratory technique, and multiple assays. Random variation (chance) leads to results being imprecise.

Repeatability is the ability of a measurement device to deliver a consistent result over several measurements under the same conditions.

Reproducibility is the ability of the device to deliver similar results under different conditions, such as a change of machine or person. Bear in mind that reproducibility cannot tell you whether the differences matter, that is a judgement call.

Unbiased

A well-designed experiment does not confound the estimate of the quantity of interest with an unwanted effect, making it impossible to estimate the quantity of interest. It controls for systematic differences between the measure and some true value, the target (see figure above). It may be difficult to achieve both accuracy and precision so there can be a trade-off between the two. Which is more important varies in different circumstances.

There are many sources and forms of bias. For example, measurement bias arises from an error in the data collection and the process of measuring. While performance bias occurs if there are systematic differences between the sample groups other than the treatment of interest.

Performance bias can be avoided by standardisation: sample groups should be in identical environments and be similar in every way apart from the applied treatments, for example by random allocation of samples to treatment groups. Both randomisation and bias will be discussed in more detail in the section Bias & Confounding Factors.

Amenable to Statistical Analysis

An investigator should never start an experiment without knowing how it is going to be analysed. They do not need to decide this alone but should seek the help of a statistician or analyst. Ideally a statistical analysis plan will be formulated, this will detail all the analyses that are to be undertaken and any transformations of the data or

changes to tests if the data collected do not meet the assumptions of the tests. As part of the statistical analysis plan the type of data to be collected (e.g. nominal[3], ordinal[4], discrete[5] or continuous[6]) should be identified (See 'Understanding Variables and Factors' in the next section).

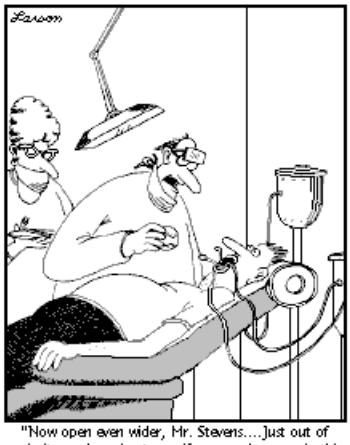
Type of Experiment

The type of experiment you want to proceed with depends on your objectives and your hypotheses. There are four main types of experiments:

- **Feasibility study:** These are pieces of research done before a main study. They are used to estimate important parameters that are needed to design the main study. For instance:
  - Standard deviation of the outcome measure, which is needed in some cases to estimate sample size
  - Willingness of participants to be randomised
  - Willingness of clinicians to recruit participants
  - Number of eligible patients
  - Characteristics of the proposed outcome measure and in some cases feasibility studies might involve designing a suitable outcome measure
  - Follow-up rates, response rates to questionnaires, adherence/compliance rates, ICCs in cluster trials, etc.

  Feasibility studies for randomised controlled trials may not themselves be randomised. Crucially, feasibility studies do not evaluate the outcome of interest; that is left to the main study. If a feasibility study is a small randomised controlled trial, it need not have a primary outcome and the usual sort of power calculation is not normally undertaken. Instead the sample size should be adequate to estimate the critical parameters (e.g. recruitment rate) to the necessary degree of precision.

- **Pilot:** A Pilot Study is a version of the main study that is run in miniature to test whether the components of the main study can all work together. It is focused on the processes of the main study, for example to ensure recruitment, randomisation, treatment, and follow-up assessments all run smoothly. It will therefore resemble the main study in many respects. In some cases this will be the first phase of the substantive study and data from the pilot phase may contribute to the final analysis; this can be referred to as an internal pilot. Alternatively at the end of the pilot study the data may be analysed and set aside, a so-called external pilot.


"Now open even wider, Mr. Stevens....Just out of curiosity, we're going to see if we can also cram in this tennis ball."

[3] Nominal: data with no ordering.
[4] Ordinal: categorical data with explicit ordering.
[5] Discrete: a finite set of (usually integer) values, often count data.
[6] Continuous: any value over an infinite range.

- **Exploratory:** exploratory studies are also called fishing trips as you go fishing in the data to see what you can find. These sorts of studies are used to generate results which go on to develop hypotheses for future experiments. Therefore, there are no hypotheses to test and they may 'work' or 'not work'.

- **Confirmatory:** These studies, also known as traditional or true experimental designs are carried out after studies of the previous kinds. They are used to test some relatively simple hypothesis stated a priori. They are to confirm hypotheses that have already been developed and not to find new ideas.

Define Factors & Develop a Good Hypothesis

In order to turn your testable question into a good hypothesis, it is important to identify and define the factors in your study.

Understanding Variables & Factors

Elements that affect your study are called factors, once measurements of these factors have been collected and are part of the data they are then known as variables (because they vary). Factors are anything that can influence the outcome of the experiment. E.g. time, weight, drug, gender, ethnic group, country, plate, cage etc.

If varying values of the factors are used they are called levels. E.g. the varying doses of each drug, male vs. female etc.

Data that are collected can be of four different types:

- **Nominal**: the data are in categories with no ordering and any observation can be assigned to only one category, this boils down to a yes/no answer if there are two levels. E.g. mouse type (wild type vs. knock out), gender, eye colour, mouse breed.
- **Ordinal**: similar to nominal, but with an implicit ordering to the categories, e.g. tumour grade, tumour stage, age category.
- **Discrete**: can take a finite range of set values, usually with an equidistance between categories, often count data, e.g. shoe size, number of cells, number of tumours.
- **Continuous**: can take any value over an infinite range, although sometimes discrete data is over a large enough range to be considered continuous, examples are weight, height, temperature.

Careful consideration should be given to what type of data to collect before the experiment is carried out, as this affects the statistical tests that can be used. It should be remembered however that discrete or continuous data can always be collapsed into categories, however, the original continuous data can never be resurrected from categories if it is not collected.

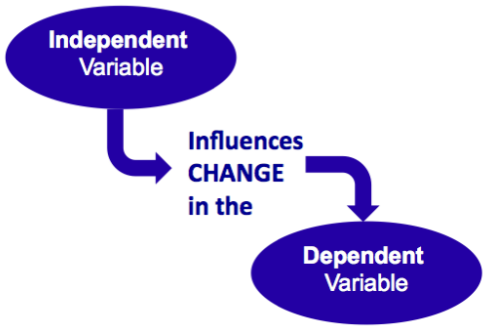
Related Observations

Most statistical tests assume an independence of observations. It is therefore important to know if any observations within the dataset are related to each other or are likely to be more similar. Are the samples or observations paired, repeated measures or matched in some way?

- **Paired:** e.g. samples taken before and after treatment in the same patient.
- **Matched:** Multiple samples from the same subject. e.g. tumour and normal samples, or matched cases and controls. Mice from the same litter, one given the drug and the other the vehicle control.
- **Repeated**: Multiple samples taken from the same cell-line. Measures taken repeatedly from the same subject.

These sorts of experimental design are more powerful than unpaired experimental design because the differences between individuals are factored out in the analysis. It is important that any pairing/matching/repeated measures that are in your data are taken into account in the analysis, other it may be underpowered and less likely to detect a treatment effect should it exist.

Independent & Dependent Variables



There are two types of variables:

- **Independent variable:** the variable that is manipulated or changed.

- **Dependent variable:** the variable that is affected by the changes made on the independent variable. This is the variable that is measured.

Developing a Good Hypothesis

A hypothesis is a tentative statement about the relationship between two or more variables. A hypothesis is a specific, testable prediction about what you expect to happen in your study. When forming a hypothesis, use your testable question as a starting point and then develop your hypothesis using the IF/THEN format. The format detail including the variables follows**:**

"IF (*the independent variable*) does this, THEN (*the dependent variable*) will result"

For example:

"IF *Smoking more than 5 cigarettes a day*, THEN *increases the risk of lung cancer*".

"IF *Interleukin (IL)-6 gene is knocked-out*, THEN *tumour growth is increased*".

The IF/THEN format is a stress-free method for writing a hypothesis and allows you to describe a causal relationship. Causality means that a set of conditions or events cause something else to change. The mission of experimental research is to illuminate this causality. Although care must be taken as causality is very difficult to prove.

Before you come up with a specific hypothesis, spend some time doing background research on your area of research. Once you have completed a literature review, start thinking of potential questions you still have. Pay attention to the discussion section in the journal articles you read. Many authors will suggest questions that still need to be explored.

Remember, a hypothesis does not have to be right. While the hypothesis predicts what the researchers expect to see, the goal of research is to determine whether this guess is right or wrong. When conducting an experiment, researchers might explore a number of different factors to determine which ones might contribute to the ultimate outcome. In many cases, researchers may find that the results of an experiment do not support the original hypothesis. When writing up these results, the researchers might suggest other options that should be explored in future studies. More formal statistical hypothesis testing will be covered in the Introduction to Statistics Course.

Choosing an Experimental Design

The Choice of Design

There are several kinds of designs you can use and the choice of design will also guide the type of data you collect and the statistical measures you use to examine your data. It is important to discuss this with a statistician. Making the correct choice will allow you to ease into the next phase of your project. The following are some general categories of research design:

- **Confirmatory or True Experimental Research Design:** This design is often thought of as a laboratory experiment and is utilized when a researcher tries to control all of the variables possible, and subjects are assigned to groups by randomisation, i.e. the independent variable is deliberately manipulated and a dependent variable is assessed.

- **Quasi-Experimental Research Design:** Quasi-experiments are very similar to true experiments but use naturally formed or pre-existing groups. For example, if we wanted to compare young and old subjects on lung capacity, it is impossible to randomly assign subjects to either the young or old group (naturally formed groups). Therefore, this cannot be a true experiment. When

one has naturally formed groups, the variable under study is a subject variable (in this case, age) as opposed to an independent variable.

- **Observational Research Design:** These studies draw inferences about the possible effect of a treatment on subjects, where the assignment of subjects into a treated group versus a control group is outside the control of the investigator (i.e. uncontrolled experimentation). For example, nutritional studies where it is very difficult to restrict what people eat (you can only test the effect of a regular intake/dose of a nutrient over and above a person's normal diet), or for ethical reasons, e.g. studies on smoking – you can not ethically ask people to take up smoking for the purposes of research, you can only work with people who have chosen to smoke. Note that it is still possible to test hypotheses in these scenarios.

- **Non-Experimental Design: T**he term is used to refer to situations in which a presumed cause and effect are identified and measured but design features such as random assignment and control groups are missing. It is worth noting that the boundaries between true experimental design and non-true experimental design are often blurry. If in doubt, you should discuss this with a statistician.

Standardise

It is essential to compare apples to apples in order to control variation and to arrive at legitimate conclusions. Controlling variation is vital as the more uniform the subjects are within a treatment group or sub-group, the fewer of them will be needed, or the greater the power of the experiment.

Standardise procedures and subjects to achieve uniformity:

- Every subject should be exposed to the same experimental conditions, differing only in the randomised treatment. e.g. a treatment must be delivered in the same way to both your control and treated groups.

- Sample groups must contain similar subjects e.g. similar age and weight of your mice in both the control wild-type and treated groups. e.g. use mice free of disease. Isogenic strains should be used if possible or mice from the same litter groups.

Minimise measurement error

Measurement error can be a big problem in experiments; if errors in measurement are too great then they can mask differences that you are trying to see. For example, to check for infection in postnatal women, the measurement of Symphysis-Fundus Distance (SFD) was measured daily by the midwives. If SFD did not fall by 1cm per day then the woman had an infection that needed treating. However, when studies on

repeatability of the measurement were carried out, it was found that the measurement error in these readings were bigger than the difference they were looking for even if the same midwife took the readings and was even worse if it was taken by different midwives.

It is important to use appropriate instrumentation to take the measurements that you are making in an experiment. Make sure that meticulous technique is applied and if possible keep things as similar as possible. For example always use the same set of scales for weighing or have the same person take the measurements. If this is not possible it is important carry out randomisation. It is especially important not to confound different ways of taking measurements with say a particular treatment group. For example do not have person A taking all the measurements on your controls and person B taking all the measurements on your treated group. Any differences between the control group and treated group will also include differences in the measurements between person A and B, the samples should be randomised to person A and B for measurement and if possible they should be blinded to treatment allocation.

If the measurement that is being taken within the experiment is particularly difficult to take or particularly prone to measurement error or lack of repeatability or reproducibility, then the measurements should be taken several times if possible. These repeated measurements could then be averaged over for analysis, or taken into account in a hierarchical model.

Experiment Controls

Planning for experiment controls is a very important part in your experiment, because it is difficult to eliminate all possible confounding variables and bias. Designing the experiment with controls in mind is often more crucial than determining the independent variable, as it increases the statistical validity of your data. There are two main types of controls:

Positive controls

Positive controls are used to check whether the procedure is effective in observing the effect and reduces the chances of false negatives. They are used to check whether the set-up is capable of producing results. For example, an established antibody is used as a positive control in ChIPSeq as it is known to produce peaks at certain genomic regions. If the control fails, then there is probably something wrong with the experiment.

Negative controls

Negative controls make sure that no confounding variable has affected the results and to take into account likely sources of bias. A sample that is not expected to work is used as a negative control and this controls for false positives. For example, a wild-type

mouse is to be compared against a gene knock-out mouse when testing the effects of inactivating a gene in vivo. A mock siRNA (scrambled sequence), is used as a negative control for silencing a gene *in vitro*.

Other forms of negative controls are sham and vehicle controls. A sham control or placebo control is used to mimic a procedure or treatment without the actual use of the procedure or test substance, e.g. putting the patient through a full surgical procedure, and making holes in their skull, but without doing anything to their brain. A sham surgery isolates incidental effects such as anaesthesia and incisional trauma. This is because it isolates the specific effects of the treatment as opposed to the incidental effects caused by anaesthesia, the incisional trauma, pre- and postoperative care, and the patient's perception of having had a regular operation.

A vehicle control is used in studies where a substance is used to deliver an experimental compound, e.g. ethanol is applied to cell lines on its own as a negative control since it's used as a vehicle for delivering the Tamoxifen drug that is being tested.

Experimental Units

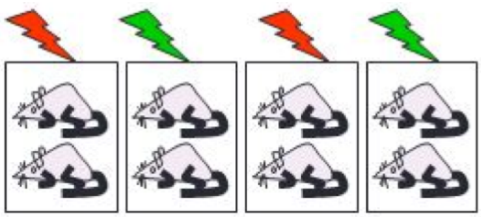
This section was taken from NCR3's Michael Festing's website: http://isogenic.info/html/6__experimental_unit.html (which no longer exists, but similar information can be found here: https://eda.nc3rs.org.uk/experimental-design-group).

Failure to correctly identify the experimental unit is a common mistake, which can result in incorrect conclusions. The experimental unit is the physical entity, which can be assigned, at random, to a treatment. In other words, it is the unit of randomisation and the unit of statistical analysis when comparing groups. Commonly it is an individual animal but this is not always the case.

Example:

The animals are housed two per cage and the treatment (see below figure: red or green) is given in the food or water.

What do you think is "N", the total number of experimental units is in this case?



Answer: two per group or four in total, since each cage is an experimental unit.

If mice in a cage are given a treatment in the diet, the cage of animals rather than the individual animal is the experimental unit. This is because mice in the same cage cannot have different treatments, and they may be more similar than mice in different cages. This means that the *p*-values in the statistical analysis may be incorrect if it is assumed that the mouse is the experimental unit. In this case the statistical analysis should normally be done using a hierarchical analysis.

Common examples of experimental units:

- The individual subject/animal
- The breeding female & litter in animal studies
- Families in human studies
- The cage or hospital or GP surgery
- Part of animal or person
- A subject for a period of time

For more examples, follow this link:
http://www.3rs-reduction.co.uk/html/3__the_experimental_unit.html

Bias & Confounding Factors

Beware the creeping cracks of bias

Bias type	Description
Selection bias	Systematic differences between baseline characteristics or treatment of the groups that are being compared.
Performance bias	Systematic differences between groups in exposure to factors other than the interventions of interest. These factors are referred to as confounding or extraneous factors.
Attrition bias	Systematic differences between groups due to samples being withdrawn from the study or excluded from the analyses.
Detection or Measurement bias	Systematic differences between groups in how outcomes are assessed or determined, such as measurement errors and inefficient use of data.
Reporting bias	Systematic differences between reported and unreported findings due to manipulation in the reporting of findings such as selective or distorted reporting, e.g. papers with more 'interesting results ' are more likely to be submitted and accepted for publication.

Bias is a factor that systematically affects the results of a study. It can be defined as the deviation of results from the truth. More specifically, it is the extent to which the statistical methods used in a study do not estimate the quantity thought to be estimated. It is the combination of any design, data, analysis, and presentation factors to produce research findings when they should not be produced or not produce a finding where they should.

Addressing the issue of bias is extremely important not only for the validity of your experiment and how you report your findings but also for the scientific community at large. In a recent commentary article in Nature, 'Beware the creeping cracks of bias' (2012), Daniel Sarewitz claimed the following:

"Evidence is mounting that research is riddled with systematic errors. Left unchecked, this could erode public trust"…

"A biased scientific result is no different from a useless one"

Early signs of bias were noted in the 1990s when researchers began to document systematic positive bias in clinical trials funded by the pharmaceutical industry. Attempts were made to reduce this problem by strict disclosure of conflicts of interest and the reporting of all clinical trials.

He mentions the now famous 2005 paper by John Ioannidis, 'Why Most Published Research Findings Are False'. Evidence of systematic positive bias appearing in research ranging from basic to clinical, and on topics ranging from genetic disease markers to testing of traditional Chinese medical practices.

However, Sarewitz claims that the problem stems from a widespread belief:

"… that progress in science means the continual production of positive findings. All involved benefit from positive results, and from the appearance of progress. Scientists are rewarded both intellectually and professionally, science administrators are empowered and the public desire for a better world is answered. The lack of incentives to report negative results, replicate experiments or recognize inconsistencies, ambiguities and uncertainties is widely appreciated — but the necessary cultural change is incredibly difficult to achieve."

He concludes that the hype from universities and journals about specific projects should be reduced and collaboration between those involved in fundamental research and those who will put the results to use in the real world, such as clinicians, should be increased.

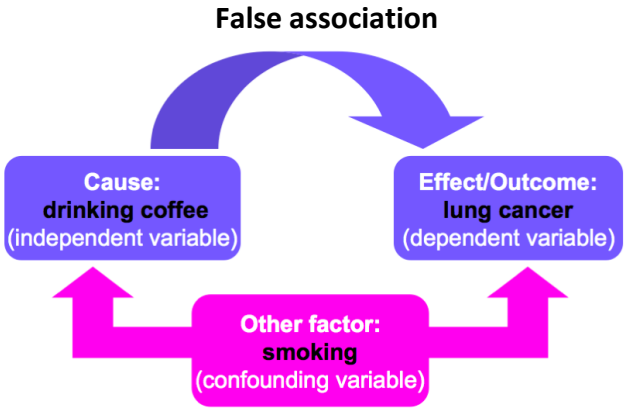
"The first step is to face up to the problem - before the cracks undermine the very foundations of science."

The second step is to ensure that all experiments are well designed.

What is a confounding factor?

When the effects of two, or more, processes on results cannot be separated, the results are said to be confounded or aliased, which is a source of bias in studies (performance bias). This type of bias may occur when there is failure to account for other variables, like age, gender, or smoking status.

The confounding factor is referred to as the third variable, as it interferes or distorts the association being studied between two other variables (the independent and dependent variables), because of a strong relationship with both of the other variables. A confounding factor is simultaneously related to the independent and dependent variable. This can mask the association between the independent and dependent variable. Making it look like there either is a relationship between them when there is none (a false positive result), or there is no relationship between them when there is one (a false negative result). It can be very difficult to tell what variables may confound an association.



A hypothetical example would be a study of coffee drinking and lung cancer. As coffee drinkers were also more likely to be cigarette smokers, and cigarette smokers are more likely to get lung cancer, the hidden effect of smoking causes the study to demonstrate that coffee drinking increases the risk of lung cancer.

If the study had collected smoking data, this could have been adjusted for in the analysis, which would remove the effect. Usually if a confounding factor is recognized, adjustments can be made in the study design or data analysis so that the factor does not confound the study results.

Other examples, this time from the media:

- ***"Democrats were less satisfied with their sex lives than Republicans". (ABC poll report).*** However, there are more democrats who are woman and woman are generally less satisfied with their sex lives. Perhaps there is a gender effect rather than a real relationship between sex life and politics.

- ***"Slightly overweight people live longer than thin people". (US Centre for disease control).*** However, the Harvard School of Public Health and the American Cancer Society later criticized the results, noting that more of the thin people were sick (and were thin because they were sick) than the overweight people. Illness was the confounding factor, illness was related to weight and illness was related to death.

Confounding factors can be obvious with hindsight but are often missed through lack of thought or money etc. Many such examples can be found in the media as media

outlets jump upon sensational results, but never pay any regard to the possibility of confounding variables. These variables can bias the results and they need to be managed and controlled, when they are controlled, then they referred to as controlled variables.

Managing confounding factors

Controlled variables are often referred to as constants, or constant variables. These are variables that are kept constant (i.e. their values are kept the same) to prevent their influence on the effect of the independent variable on the dependent variable. It is important to ensure that these variables are isolated, because an error may occur if an unknown factor influences the dependent variable. This could lead to the null hypothesis being correctly rejected, but for the wrong reason.

Error	Definition
Type I	Incorrectly reject the null hypothesis when it is true
Type II	Failure to reject the null hypothesis when it is false
Type III	The null hypothesis is correctly rejected, but for the wrong reason

Inadequate monitoring of controlled variables is one of the most common causes of researchers incorrectly assuming that a correlation or association leads to causality. Designing the experiment with controls in mind is often more crucial than determining the independent variable. Poor controls can lead to confounding variables, and will damage the internal validity of the experiment and waste time and resources.

Effective monitoring of controlled variables to reduce bias can be achieved by:

- Correct selection of experimental units.
- Randomisation of the experimental units.
- Randomisation of the order in which measurements are made.
- "Blinding" and the use of coded samples.

Randomisation

Randomisation ensures that each experimental unit has a known probability of receiving a particular treatment (in most cases this is 50:50, but it does not have to be). It reduces selection bias, increases the validity of the findings and, in principle, is always an appropriate and desirable aspect of good experimental design when two or more treatments are compared. Randomisation balances known and unknown variables between groups and therefore reduces the effect of confounding variables as they should by chance have a similar distribution between each group.

Randomisation should also extend to cage placement within rooms in the animal house and the order in which experimental treatments and assessments of the animals/cages are made. Anywhere an arbitrary decision is made in an experiment it should be done via randomisation.

Appropriate methods should be used for randomisation, use a random number table or generator program as even tossing a coin may be subject to bias. To avoid bias someone other than the experimenter should carry out randomisation, often this is a job for the statistician. Many methods may incorrectly be called random but they are not truly random and are therefore, subject to bias. These include experimenter choice, alternating allocation, date of entry, date of birth, hospital/sample number, etc etc.

Randomised Block Design

It is not sufficient for good experimental design to just randomly place samples on plates (i.e. a "completely randomised" design), it is key that the randomisation is controlled so that the experimental units are almost perfectly balanced (i.e. a "randomised block design").

Randomised block designs are where experimental units, for example, animals are first divided into homogeneous groups before the groups are randomly assigned to a treatment group. They can be used to introduce variation in the groups of animals (e.g. sex, age, severity of disease) in a controlled way without the need for larger numbers of animals.
The following example, a case-control Parkinson's Disease genotyping study, can be used to illustrate the general concept of block randomisation.

Example plate layout for a ~4,000-subject Parkinson's Disease genotyping study:

(source: http://blog.goldenhelix.com/?p=322)

The study had 2,000 cases and 2,000 controls collected from four different sites using three DNA extraction methods. This looks like an experimental nightmare but is actually not difficult to manage. The following table illustrates a block randomisation involving case/control status, site, and DNA extraction method. Each row of the table contains counts of the various experimental units to be randomised across plates.

The case/control status is the most important variable to randomise across plates. With quantitative traits (continuous variable), some form of discretisation (categorisation) into experimental units can be employed. While randomised plating will not remove the confounding effect due to the non-ideal data collection of different numbers of cases and controls by site and DNA extraction method, we can at least ensure these will not be further confounded with plate artefacts. So if we need to correct for data distortions due to site or DNA kit later, this can be done using appropriate statistical methods.

Block	Case Status		Site				DNA Extraction Method			Number
	Case	Control	1	2	3	4	1	2	3	
1	X		X				X			407
2	X		X					X		42
3	X		X						X	61
4	X			X				X		854
5	X				X			X		417
6	X					X		X		219
7		X	X				X			191
8		X	X					X		684
9		X	x						x	28
10		X		X				X		684
11		X			X			X		300
12		x				x		X		113

How does this table translate into plating the samples?

As there are 4000 samples a total of 45 plates each plate having 96 wells or positions (a total of 4320 wells or positions), are required. Only one sample can go into one well/position. The remaining 320 wells across the plates were reserved for male and female control samples and other controls.

The 4000 samples can be split into 12 homogenous groups (blocks) based on common baseline characteristics (Case status; Site; and DNA Extraction Method). The 407 samples from Block 1, containing cases from Site 1 with DNA Extraction Method 1, would be evenly divided at random among the 45 plates, resulting in either nine or ten samples per plate. So there are three steps here:

- Calculate how many subjects in that EU there should be on each plate (e.g. for EU 1, 407/45 = 9.04444).

- Randomly assign that number of subjects (e.g. 9 or 10 for EU 1) to the 45 plates.

- Randomly assign each subject to the 96 different plate positions.

We similarly divide Experimental Units 2 through to 12 to construct 45 plates with 90 samples or less using the same three steps for each experimental unit.

Note that it is not sufficient for good experimental design to just randomly place cases and controls on plates (i.e. a “completely randomised” design). It is key that the randomisation is controlled so that the experimental units are almost perfectly balanced (i.e. a “randomised block design”) where possible. What we want is a balance in the number of cases and controls to plate, which is important if we are expecting plate effects or if a plate drops out. Studies that place samples completely at random

can, by chance alone, have many plates being quite unbalanced and this can create unwanted, spurious plate-driven associations, or cause problems if a plate is lost completely. Bioinformatics are available to help you with the plate layout for your experiment.

Blinding

"*When humans have to make observations there is always the possibility of bias*". Cochrane et al (1972).

Blinding is the deliberate withholding of treatment allocation to avoid bias. Blinded assessment, where appropriate, minimises any bias (performance and detection bias) in the qualitative scoring of subjective experimental observations, improving the rigour of the experimental method and the scientific validity of the results obtained. If you do not blind, then you know, as the experimenter, which animals or subject had which intervention. So you might allow that knowledge, even unconsciously, to affect close calls on measurements you take, or cause you to retake a suspicious measurement.

Wherever possible use a study number as a code to blind everyone to the treatment. This is particularly important when making measurements, scoring histological sections or measuring behaviour. Blinding may be difficult in some cases such as when comparing two mouse strains, which differ in coat colour. In situations where blinding is impossible, reliance needs to be placed in randomisation of the order in which the animals are tested.

There is strong evidence of a placebo effect with medicine, where, if people believe that they are receiving a medicine, they show some signs of improvement in health. A blind experiment reduces the risk of bias from this effect, giving an honest baseline for the research, and allowing a realistic statistical comparison. Researcher or clinician blinding may still be needed in addition to patient blinding.

Reviews of animal research in the field of emergency medicine found that studies which did not use randomisation and blinding to reduce bias, when comparing two or more experimental groups, were significantly more likely to find a difference between the treatment groups (Bebarta et al, 2003; Macleod et al, 2008).

Experiments can be not blinded, single, double or triple blind, depending on who out of the subject, researcher and analyst/statistician are blinded to group allocation. It is not always possible to blind all three, but it is important that an experiment reaches its potential degree of blinding, so that everyone who can be blinded, is blinded.

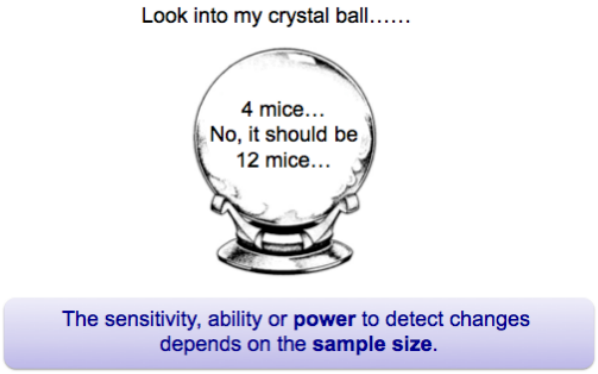
Replication, Sample Size & Power

How many replicates?

The **number of replicates** or the size of the experiment is an important aspect of any experiment but is one that is frequently overlooked until the last minute. Although there is no one simple method to work out how many replicates you will need, there are some factors you must bear in mind.

Some other names for the number of replicates:

- Sample size
- Total N
- Number of participants
- Number of experimental units
- Number of observations
- Group size
- Experiment size



These names are used interchangeably, although they have slightly different interpretations. For example, group size is the size of an individual group, whereas total *N* is the overall size of the experiment (the sum of the individual group sizes). When discussing the number of replicates it is important to be clear if you are referring to the total number of replicates overall or the number per group.

There are many factors that determine the number of replicates that you should use. Preferably you should use a formal sample size calculation (discussed later in this section), but other relevant factors are:

- Resources available (e.g. finances, time and amount of material)
- Experimental goals
- Technology reliability
- Numbers of samples that fit on a microarray/plate/sequencing run

Be careful when making money the limiting factor: it would be better not to waste money on an experiment that will be too small to show a meaningful difference at the end of it.

The number of replicates you need is heavily dependent on the size of effect that you want to detect (e.g. log-fold change of 2, difference in means of 0.5). The smaller the difference that you wish to detect, the harder it will be to find, and therefore the more replicates that will be needed.

The number of replicates you use must be large enough to be representative of the population that the results will be generalised to. For example, if your population contains five different breeds of mice, there is no point having only four replicates as at least one of the breeds will not be represented.

The number of replicates must be sufficient to ensure that biologically or clinically meaningful results are likely to be detected. A formal sample size or power calculation

can be used to bring statistical significance in line with biological or clinical significance. Otherwise your experiment might achieve a statistically significant result which is not biologically or clinically meaningful, or vice versa.

The number of replicates must be large enough to model the noise levels or variability in the system. The noisier the system you are studying, the harder it will be to detect the signal from the noise and give meaningful results. For this reason it is important to try and control as much within group variability as possible by keeping as much the same between replicates as possible. Then the within group variability is less likely to drown out the between group variability.

Animal and patient samples tend to be more variable than cell line samples and therefore require larger sample sizes. Tumour samples are often heterogeneous so tissue samples from tumours often contain mixtures of different tissue types, there may be contamination from normal tissue for example, this makes tumour samples more variable between patients/ animals and so this needs to be accounted for with a larger sample size. Its not just the tumours themselves that are variable, there will be variability in the patients and animals that are included in the study, they may be different ages, different sexes, some may have different concomitant diseases that all add to the patient variability and make it harder to see differences between groups. A larger sample size will allow a more complex analysis that can account for these factors and make it easier to detect the signal within the noise.

In many studies not all samples/subjects make it through to the end of the study, they 'drop-out', this phenomenon is known as sample attrition. There can be many reasons for this attrition. In patient or animal studies, subjects may die before the end of the study, or be too ill to continue participating or may be withdraw consent to continue in the study. Samples can be of too poor quality to sequence or not provide enough material to sequence, samples can be lost or vials dropped. The number of replicates must allow for sample attrition. All experiments should be robust to a small amount of sample attrition, so that losing one sample from your experiment does not prevent you from getting useful results. However, some studies are particularly prone to sample drop-out. It is important to know why subjects or samples dropout of a study, because the dropout can be related to the study endpoint. In this case subjects or samples cannot be ignored in the analysis. If observations are missing in a non-random fashion then ignoring them in the analysis will bias your results. Imagine in a treatment versus control drug study, half of the patients drop out of the treatment arm due to the side effects of the treatment, you will get very different results if you analyse half or all of the patients in the treatment arm.

You can maximise the use of the replicates you have by keeping variability within a group to a minimum. Wherever possible you should try and keep all factors the same within each group, including reagents, person handling samples, time exposed to treatment, etc. In an ideal experiment, the only variability is the randomised difference that you are trying to detect. In the real world this is not possible, but we should aim to be as close to this ideal experiment as possible.

Larger sample sizes are needed when…

- **You have a large number of uncontrolled variables which interact unpredictably**: A larger number of samples are required when there are large numbers of variables that it is not possible to control and there is no way of knowing how they interact (or what effect) they have on each other. In experiments such as this it will be much harder to determine what is going on so a more complex design and/or analysis will be required and, therefore, more samples may be needed.

- **The total sample set is going to be analysed as several sub-sets:** in this case each subgroup, rather than just the overall sample, will need to be large enough to detect the effect size of interest. For example, you may wish to look at treatment effects separately in males and females, therefore the male and female subgroups both need to be sufficiently large to detect the effect size of interest if it exists. If possible this a factorial design should be employed so that the overall number is kept to a minimum.

- **When the population you are studying has many variables:** then a larger sample size is required to allow for multiple hypothesis testing of all the variables (see earlier section on Factorial Design). If there are many characteristics varying within the population then it can be difficult to disentangle what is happening in the experiment, therefore more subjects are required and more complex statistical analysis will be required. Interaction effects between variables can be very small and therefore, very large sample sizes are required to detect them.

- **You want to detect small effect sizes:** The smaller the difference to be detected the harder it will be to find; therefore larger samples sizes will be required. It is important to have some idea of the effect sizes of interest before carrying out a study. This might be for example: difference in means, differences in survival times or log fold changes.

- **You expect samples or subjects to be lost by attrition**: If high attrition of subjects or samples is expected, that is low number of subjects or samples making it through to the end of the experiment or study, larger sample sizes are required to ensure that there are sufficient subjects or samples to analyse at the end of the study or experiment. For example, samples may drop out of an experiment because they are of poor quality, such as having low RIN values in microarray experiments, there may be problems with positive or negative control in part of an experiment leading to sample dropout, or in a clinical study patients may be too ill to complete the entire study. It is important to know why subjects or samples drop out of a study. If the dropout is related to the study endpoint, then these subjects or samples cannot be simply ignored in the analysis. If observations are not missing at random then ignoring them in the analysis can bias the results.

What is Power?

Power is the flip side of significance but is often overlooked. A significant result is the aim of any experiment, and in order to achieve a reproducible significant result, the experiment must have sufficient power. Power is the probability of detecting a specified difference, if it exists, within the population. The difference of interest depends on the experiment, for example: a log fold change in a microarray experiment, a difference in survival times in a clinical study, or a change in the size of a tumour in a mouse study.

If all other parameters remain the same, a larger experiment will always have more power than a smaller experiment. However, if an experiment is too large and a smaller experiment would have achieved the same statistical result, it is overpowered and it has wasted subjects, money, time and effort, and is potentially unethical if animals or patients have been used. On the other hand, if an experiment is too small, it may lack power and miss important differences that do actually exist. Therefore, an underpowered study also wastes resources and can be unethical. It is important to know what effect size is important and to carry out a sample size calculation to ensure that your experiment is sufficiently powered.

Sample Size Calculations

Sample size calculations, power calculations and power analysis (the terms are used interchangeably) are a way of determining the appropriate number of replicates (the sample size) for a study. There are many different forms of sample size calculations to suit different types of statistical test, as different statistical tests have different parameters their respective sample size calculations have different parameters.

Most parameters fall into six factors (listed below) which are intimately linked, so if we know five of them we can estimate the sixth one. In practice the final three tend to be fixed for a study and we are usually only estimating one of the first three.

- **Power** (ω or $1 - \beta$): the probability of detecting the specified effect size given that it exists. If a sample size calculation is carried out then power is often fixed at 80% or sometimes 90% in a more stringent sample size calculation. For a power calculation, we calculate the power for a given sample size. As power is a probability (converted to a percentage) it may take on any value between 0% and 100%.

- **Sample size** (n or N): which is usually what is being estimated, depends on the type of power calculation used. In general it is the number of observations per group, or it could be the total number of observations in the whole experiment (For example in survival studies it is the number of deaths or other events, for sensitivity it is the number of patients with the disease). Sample size calculations will usually not produce an exact integer value e.g. 213, 579, 865, so to achieve sufficient power sample sizes are always rounded up to the nearest integer.

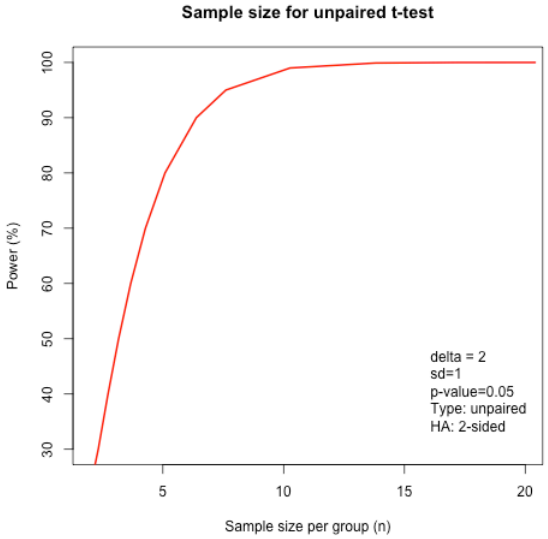
- **Effect size** (δ): The size of the difference of interest depends on the nature of the sample size calculation: for example, it could be the difference in means, the difference in survival course, or a difference in proportions. The size of the difference that is of interest is totally dependent on the clinical or biological situation and may be different for each study. The difference should be either biologically or clinically meaningful.

- **Variability** (sd): measure of the amount of noise in the experiment. The exact measure of variability required will depend on the sample size calculation being carried out. For an unpaired t-test, it is the pooled standard deviation from the two groups, for the paired t-test it is the standard deviation of the differences and as will be seen later, for microarray experiments it is the vector of standard deviations of each of the probes in the control group. A pilot study is often required to estimate the standard deviation to use in the sample size calculations. Pilot studies are often carried out only on the control group and it is assumed that the standard deviation will be similar in the treatment group. However, the standard deviation is difficult to estimate and pilot studies may provide a poor estimation of the variability due to their small sample size. Where possible, an estimate from a large study from the literature may be a better estimate. Where a pilot study has been carried out it is important that this is separate from the main study and that samples are not included in both the pilot and the main study, otherwise bias may be introduced.

- **The significance level** (α): is the cut-off in the p-value that will be taken to be significant at the end of the experiment. The usual cut-off is 0.05, that is one in twenty tests will be significant just by chance, but it is not possible to tell which tests are significant just by chance. A more stringent study might take 0.01 as a cut-off, i.e. one in a hundred tests will be significant just by chance.

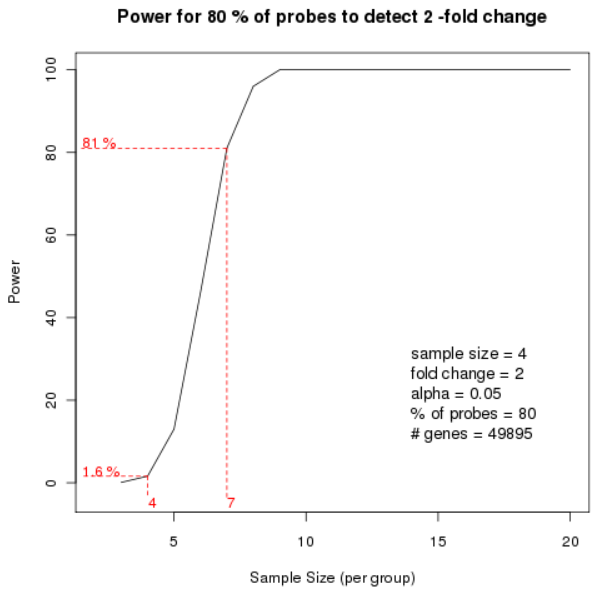
- **The alternative hypothesis** (H_A or H_1): whether a change in one (pre-specified) direction or a change in both directions is of interest. For example, if the null hypothesis is that the difference in median survival time between treatment and control groups is 2 years. A one-sided alternative hypothesis could be that the median survival time in the treatment group is greater than 2 years longer than the control group. A two-sided alternative hypothesis could be that the difference in median survival time between the treatment and control group is not equal to 2 years. Another example for microarray gene expression: either up or down gene regulation (one-sided) or both up and down gene regulation (two-sided).

Sample Size Calculation Example 1: unpaired t-test



In a study mice are randomised to a treatment and a control group. At the end of the study a 2g difference in the intake of food is of interest. From a pilot study it is found that the standard deviation of weight of food intake is 1. In this case a pilot experiment was carried out on just the control group to estimate the standard deviation. It was then assumed that the standard deviation would be the same in the treatment group. The investigators wish to power their study so that a 2g difference in means is likely to be significant at the 0.05 significance level, using a two-sided test. The power curve above is plotted for this experiment. The curve plateaus at around 10 samples per group, therefore, assuming that the parameters are an accurate reflection of the situation, there is little benefit in increasing the sample size beyond 10 mice per group. At least 80% power is estimated to detect a mean difference of 2 and this is achieved with 6 mice per group and 90% power with 7 mice per group. In this example experiment, 6 or 7 mice per group would be a reasonable amount to use.

Sample Size Calculation Example 2: microarray experiments



Sample size calculations for microarray experiments are slightly different to usual sample size calculations. For most sample size calculations there is usually one primary outcome of interest, or at most a handful. In microarray experiments we are carrying out calculations for thousands of hypothesis tests, one per probe (which may mean multiple tests per gene). Therefore, we may not require all of our hypothesis tests to have sufficient power. In this case, the sample size calculations contain an additional parameter: the percentage of probes achieving power.

In the example plot above, we are interested in a list of differentially expressed genes (probes) that are defined by a fold change of 2 between the control and a treatment

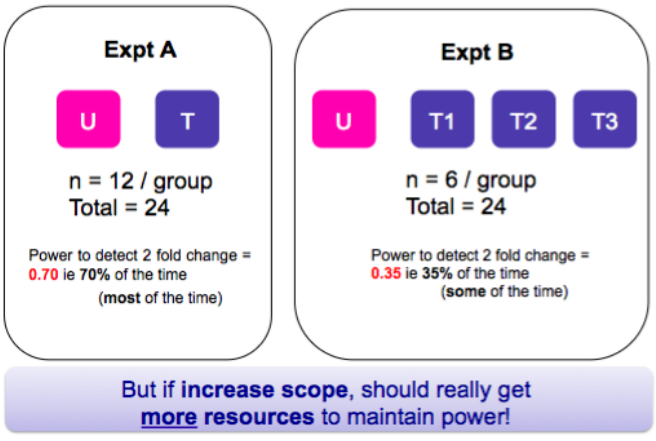
group and a p-value of 0.05. The example is based on pilot data from a control group with 4 replicates and there are 49895 probes on the microarray chip being used. As the standard deviations used for the calculation are only from the control group, it is assumed that the standard deviations for each probe are similar in the treatment group. It is also assumed that we can obtain a good estimate of the standard deviations of each of the probes measured in the control group with only 4 replicates.

In this example only 80% of the probes (genes) are required to achieve power. Looking at the plot of sample size per group against power, if only 4 replicates were used, as was the case for the pilot, then only 1.6% power would be achieved by at least 80% of the probes. A sample size of 7 replicates per group would be required for at least 80% of probes to achieve at least 80% power.

Looking at the plot, it plateaus at 100% power, for around 10 replicates per group. If only 80% of probes are required to achieve sufficient power then there is little benefit in increasing the number of replicates beyond 10, as it is not possible to gain an increase in power to detect a 2-fold change, with a p-value of 0.05, in 80% of the probes. Of course by increasing the number of replicates beyond 10, we could increase the power to detect a fold change of less than 2 and/or with a more stringent p-value and/or in more than 80% of the probes. When carrying out sample size calculations it is worth changing the parameters to see what effect this has on the sample size calculations to come up with the most optimal sample size for your particular experiment.

Which experiment: increase or reduce scope?

Including more treatment groups without increasing the number of samples will significantly lower the power of the experiment.



In experiments A and B (see figure) we are looking for a fold change of 2, with the same standard deviation of 0.5, with a two-sided significance level of 0.05. The only difference is experiment A has just one treated group where experiment B has three treated groups. Both experiments have just one untreated (control) group.

If we only have 24 samples available to use in this experiment, then in experiment A, we have 12 samples per group, giving us 70% power to detect a fold change of 2. That is to say we should see this biologically meaningful result most of the time (if it exists). If, however, we have to split our resources between four groups rather than two, we have 6 samples per group and the power to detect a fold change of 2 halves to 35%. In other words, we will only see the biologically meaningful result some of the time (if it exists).

By including more treatment groups we have significantly lowered the power of the experiment, from 70%, just about reasonable, to 35% where it would be questionable whether the experiment is worth carrying out.

So what are our options? We can either decrease the scope of the experiment, that is having fewer treatment groups which has the effect of increasing the number of subjects per group and raising the power of the experiment, or we can increase the scope of the experiment, which has the effect of decreasing the number of subjects per group and lowering the power of the experiment. We can either focus in on one treatment comparison, which has the advantage that we will be able to estimate that effect well but might need further experiments to answer all of the questions, or we can expand the scope of our experiment to include several treatments and perhaps not answer the question so well if the number of subjects is limited. However, we will be using the same control group for many comparisons.

If we are increasing the scope of an experiment then more resources are required in order to maintain the power of an experiment. If sufficient subjects are available it can be beneficial to test several treatments simultaneously as the control group can be used as the comparator for each treatment group if all can be run concurrently. This can reduce the number of control subjects required and allow direct comparison of the treatment groups as well. If treatment is for a chronic condition then a cross-over design can increase the power, in this case each subject has two or more treatments in a randomised order and treatment comparisons are within patients which can also increase the power of the experiment. However, there can be carry over effects, that is where the effects of a treatment persist to the next treatment period, which is why the order of treatments should be randomised.

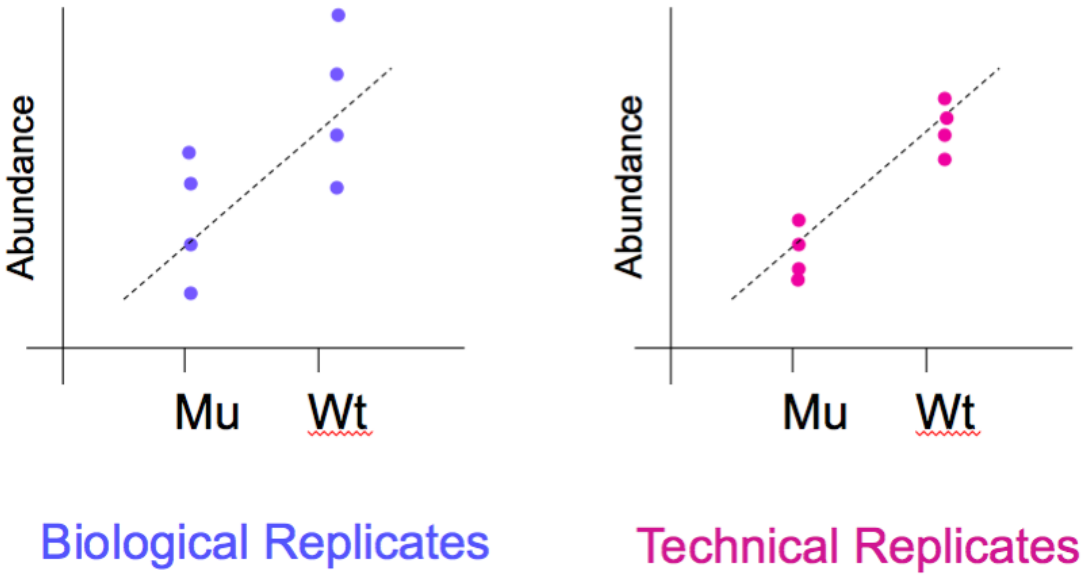
Biological or Technical Replicates?

It is important to be clear on whether you are collecting biological replicates or technical replicates, or both, as the two types of replicates are very different:

Replicate: an independent observation obtained under conditions as identical to the original as the nature of the investigation will permit.

Technical replicates: measure a quantity from a single source. The differences are based only on technical issues in the measurement. For example, if I weigh the same mouse three times, do I get different weights and how different are the measurements? This is an example of repeatability, the closeness of the results obtained in the same test material by the same observer or technician using the same equipment, apparatus and/or reagents over reasonably short intervals of time.

Biological replicates: measure a quantity from different sources under the same conditions, e.g. Tumours from 5 different patients with lung cancer may show similar gene expression patterns. These replicates are useful to show what is similar in your replicates and how they are different from a different set of conditions (e.g. treated with a drug, or normal lung).



Running both biological and technical replicates as part of the same experiment adds unnecessary complexity to your data analysis and, unless handled correctly, you risk pseudoreplication[7] (which is discussed in length by Lazic et al, 2010).

For example, you can consider setting up experiments that comprise multiple technical replicates of one biological sample if want to measure the technical variation of your system. You could then compare any changes observed between biological replicates against the background of technical variation to report the reliability of the biological differences.

Your technical variation should be considerably lower than your biological variation, as illustrated in the example above for the gene expression abundance of: mutant (Mu) and wild type (Wt) mice. On the left panel are four biological replicates, i.e. samples from four different mice for both Mu and Wt sample groups. On the right panel are four technical replicates from the same individual mutant and the same individual wild-type mice. The scale on the y-axis is the same in both figures.

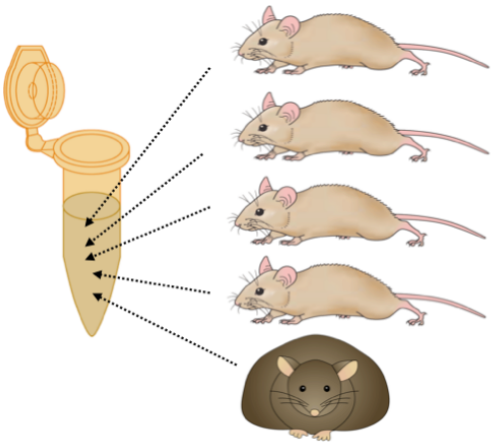
Technical replicates should not be included in your experiment unless you are optimising and testing a technical procedure. If you are interested in biological effects, technical replicates will not increase the power of your experiment. Only biological replicates will increase the power of your experiment. Do not include technical replicates unless you have an analysis plan for them. It would be a waste of resources to include controls on every plate in a large-scale experiment and then do nothing with them.

There is much debate on what constitutes a biological replicate in the case of cell line samples. Samples from different cell lines that originate from different patients are

[7] Pseudoreplication typically occurs when the number of observations or the number of data points are treated inappropriately as independent replicates. Observations may not be independent if (1) repeated measurements are taken on the same subject, (2) the data have a hierarchical structure (e.g. in cell culture experiments), (3) observations are correlated in time, or (4) observations are correlated in space.

biological replicates. Whether different passages of a cell line or the same passage of cells grown separately are biological replicates is a grey area. If you have only used one cell line in your experiment then the results will only apply to that cell line and therefore, may be due to quirks in that cell line. If you include different cell lines then your results are more generalisable.

Pooling Replicates



It is tempting to pool samples to save on processing costs. Maybe for reasons of practical necessity you need to pool because you are working with very small amounts of material. There is considerable disagreement among practitioners and statisticians about whether to pool individual samples. For instance, with global gene expression analysis, in theory if the variation of a gene among different individuals is approximately normally distributed, then pooling *N* independent samples would result in reduction of variance.

In principle we could then further reduce the variation by making replicates of the pool. Since technical variation is usually less than individual variation, this strategy would in theory give us more accurate estimates of the group means for each gene. The idea behind this motivation is that differences due to subject-to-subject variation will be minimised, making large effects easier to find. This is often desirable when primary interest is not on the individual (e.g. making a prognosis or diagnosis), but rather on characteristics of the population from which certain individuals are obtained (e.g. identifying biomarkers or expression patterns common across individuals).

In practice the distribution of expression levels of many genes among individuals are not roughly normal; often there are more very high values (outliers) than the normal distribution. Some individual samples have levels of stress response proteins and immunoglobulins five to ten fold higher than typical. This can be due to many factors unrelated to the experimental treatment: for example, individual animals or subjects may be infected, or some tissue samples may be anoxic (a total depletion of oxygen) for long periods before preservation, which allows cells to respond to stress (Prichard et al, 2002). It is easier to detect this, if individual samples are processed and not pooled.

In some studies (Terry Speed, unpublished data; Kendziorski et al, 2005), where the same samples were analysed by pooled and unpooled designs, the majority of genes that were identified as differentially expressed between two groups, turn out to be extreme in only one individual. Also, if one pools samples, there is no way to estimate variation between individuals, which is sometimes important and often interesting. Without biological replication, outliers cannot be found and appropriate variance components cannot be estimated.

A Final Note

Whatever experiment you are planning it is advisable to seek the advice of the Genomics and Bioinformatics Cores at one of the Experimental Design meetings. They will be able to advise you whether your design is suitable to answer the scientific question(s) that you wish to answer, whether the number of replicates that you are proposing is reasonable to answer your question(s) and whether any confounding factors may get in the way of answering your question(s). They can also advise if any aspects of your design require randomisation and provide a randomisation schedule.

Good luck with your experiments!

References

Bebarta V, Luyten D, Heard K (2003) Emergency medicine animal research: does use of randomisation and blinding affect the results? Academic Emergency Medicine 10(6): 684–7.

Cochrane AL (1972) The History of the Measurement of III health. International Journal of Epidemiology 1: 89–92.

Fisher RA (1951). Design of Experiments, 6th ed. (Edinburgh, UK: Oliver and Boyd).

Fisher RA (1960). Design of Experiments. (New York: Hafner Publishing Company, Inc).

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

Kendziorski et al (2005). On the utility of pooling biological samples in microarray experiments. PNAS, 102(12):4252-7

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG (2010) Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. PLoS Biol 8(6): e1000412. doi:10.1371/journal.pbio.1000412

Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PloS ONE. 2009;4:e7824.

Lazic SE (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci* 11:5.

Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, et al. (2008) Evidence for the efficacy of NXY -059 in Experimental Focal Cerebral Ischemia is confounded by study quality. Stroke 39: 2824–2829.

NC3Rs, (2006), Why do a pilot study? http://www.nc3rs.org.uk/downloaddoc.asp?id=400

Niewenhuis et al. (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neurosci. 14:1105

Pritchard et al (2002) Project normal: defining normal variance in mouse gene expression. PNAS, 98(23):13266-71

Richter SH, Garner JP, Würbel H , (2009), Environmental standardization: cure or cause of poor reproducibility in animal experiments?, Nature Methods, 6, pp 257-261

Sarewitz, D (2012) Beware the creeping cracks of bias. *Nature* 485, 149

Tung-Tien Sun (2004). Excessive trust in authorities and its influence on experimental design. *Nature Reviews Molecular Cell Biology*

Index